Read M&M §7.1 and §7.2, ignoring starred parts. Reread M&M §3.2. The effects of estimated variances on normal approximations. t-distributions. Comparison of two means: pooling of estimates of variances, or paired observations.

In Lecture 6, when discussing comparison of two Binomial proportions, I was content to estimate unknown variances when calculating statistics that were to be treated as approximately normally distributed. You might have worried about the effect of variability of the estimate. W. S. Gosset ("Student") considered a similar problem in a very famous 1908 paper, where the role of Student's t-distribution was first recognized.

Gosset discovered that the effect of estimated variances could be described exactly in a simplified problem where n independent observations $X_1, \ldots, X_n$ are taken from a normal distribution, $N(\mu, \sigma)$. The sample mean, $\overline{X} = (X_1 + \ldots + X_n)/n$ has a $N(\mu, \sigma/\sqrt{n})$ distribution. The random variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution. If we estimate σ^2 by the sample variance, $s^2 = \sum_i (X_i - \overline{X})^2/(n-1)$, then the resulting statistic,

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

no longer has a normal distribution. It has a ***t-distribution on $n-1$ degrees of freedom***.

> REMARK.   I have written T, instead of the t used by M&M page 505. I find it causes confusion that t refers to both the name of the statistic and the name of its distribution.

As you will soon see, the estimation of the variance has the effect of spreading out the distribution a little beyond what it would be if σ were used. The effect is more pronounced at small sample sizes. With larger n, the difference between the normal and the t-distribution is really not very important. It is much more important that you understand the logic behind the interpretation of Z than the mostly small modifications needed for the t-distribution.

1.   How different is the t distribution from the normal?

In each of the following nine pictures I have plotted the quantiles of various t-distributions (with degrees of freedom as noted on the vertical axis) against the corresponding quantiles of the standard normal distribution along the horizontal axis. The plots are the population analogs of the normal (probability or quantile—whichever name you prefer) plots from M&M Chapter 1. Indeed, you could think of these pictures as normal plots for immense sample sizes from the t-distributions.

The upward curve at the right-hand end and the downward curve at the left-hand end tell us that the extreme quantiles for the t-distributions lie further from zero than the corresponding standard normal quantiles. The tails of the t-distributions decrease more slowly than the tails of the normal distribution (compare with M&M page 506). The t-distributions are more spread out than the normal.

The spreading effect is huge for 1 degree of freedom, as shown by the first plot in the first row, but you should not be too alarmed. It is, after all, a rather dubious proposition that one can get a reasonable measure of spread from only two observations on a normal distribution.
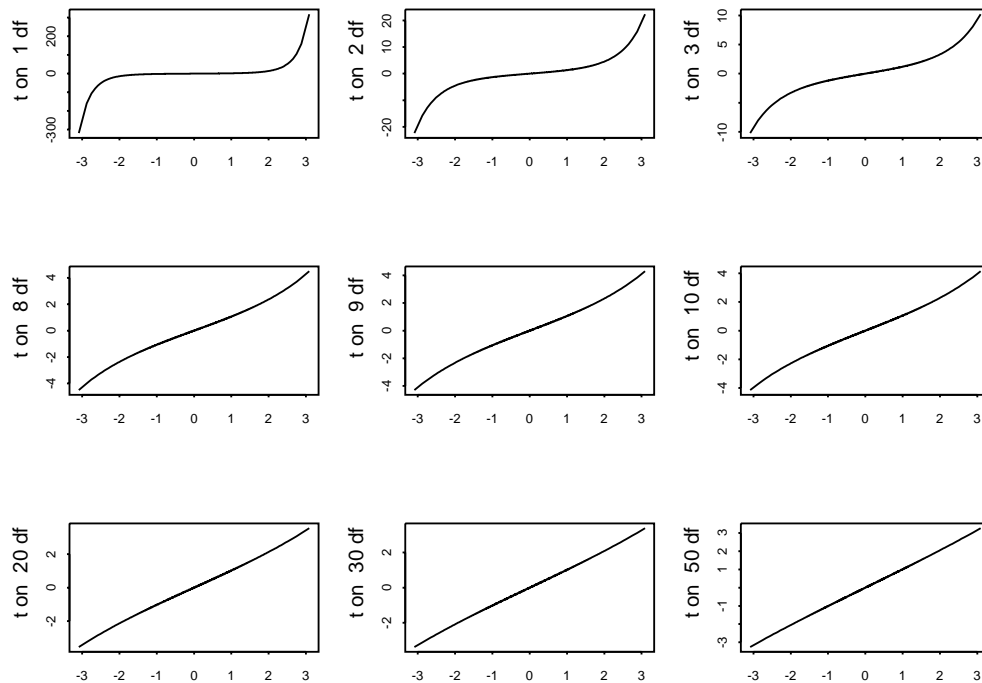
With 8 degrees of freedom (a sample size of 9) the extra spread is noticeable only for normal quantiles beyond 2. That is, the differences only start to stand out when we consider the extreme 2.5% of the tail probability in either direction.

The plots in the last row are practically straight lines. You are pretty safe (unless you are interested in the extreme tails) in ignoring the difference between the standard normal and the t-distribution when the degrees of freedom are even moderately large.

The small differences between normal and t-distributions are perhaps easier to see in the next figure.
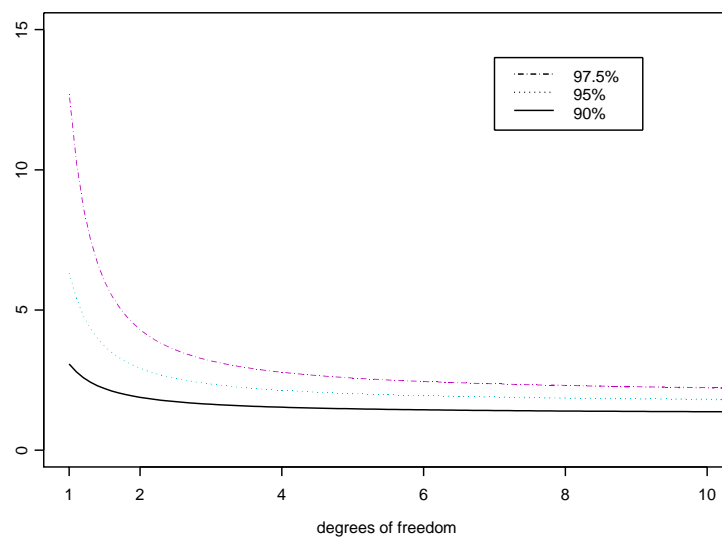
Quantiles for the t-distribution, plotted against degrees of freedom. Corresponding standard normal values shown as tickmarks on the right margin.

The top curve plots the quantile $q_n(.975)$, that is value for which

$$\mathbb{P}\{T \le q_n(.975)\} = 0.975 \qquad \text{when } T \text{ has a t-distibution with } n \text{ degrees of freedom,}$$

against n. The middle curve shows the corresponding $q_n(0.95)$ values, and the bottom curve the $q_n(0.90)$ values. The three small marks at the right-hand edge of the picture give the corresponding values for the standard normal distribution. For example, the top mark sits at 1.96, because only 2.5% of the normal distribution lies further than 1.96 standard deviations out in the right tail.

Notice how close the values for the t-distributions are to the values for the normal, with even only 4 or five degrees of freedom.

> REMARK.   You might wonder why the curves in the previous figure are shown
> as continuous lines, and not just as a series of points for degrees of freedom
> $n = 1, 2, 3, \ldots$. The answer is that the t-distribution can be defined for non-integer
> values of n. You should ignore this Remark unless you start pondering the formula
> displayed near the top of M&M page 549. The approximation requires fractional
> degrees of freedom.

<1>   **Example.**   From a sample of n independent observations on the $N(\mu, \sigma)$ distribution, with known σ, the interval $\overline{X} \pm 1.96\sigma/\sqrt{n}$ has probability 0.95 of containing the unknown μ; it is a 95% confidence interval.

If σ is unknown, it can estimated by the sample standard deviation s. The 95% confidence interval has almost the same form as before, $\overline{X} \pm Cs/\sqrt{n}$, where the unknown standard deviation σ is replaced by the estimated standard deviation, and the constant 1.96 is replaced by a larger constant C, which depends on the sample size. For example, from Minitab (or from Table D at the back of M&M), $C = 12.71$ if $n - 1 = 1$, or $C = 2.23$ where $n - 1 = 10$. If n is large, $C \approx 1.96$. The values of C correspond to the top curve in the figure showing tail probabilities for the t-distribution.

2.   How important is the t distribution?

In their book "Data Analysis and Regression", two highly eminent statisticians, Frederick Mosteller and John Tukey, pointed out that the "value of Student's work lay not in a great numerical change" but rather in

- recognition that one could, if appropriate assumptions held, make allowances for the "uncertainties" of small samples, not only in Student's original problem but in others as well;
- provision of numerical assessment of how small the necessary numerical adjustments of confidence points were in Student's problem and . . . how they depended on the extremeness of the probabilities involved;
- presentation of tables that could be used . . . to assess the uncertainty associated with even very small samples.

They also noted some drawbacks, such as

- it made it too easy to neglect the proviso "if appropriate assumptions held";
- it overemphasized the "exactness" of Student's solution for his idealized problem;
- it helped to divert attention of theoretical statisticians to the development of "exact" ways of treating other problems.

(Quotes taken from pages 5 and 6 of the book.) Mosteller and Tukey also noted one more specific disadvantage, as well as making a number of other more general observations about the true value of Student's work.

In short, the use of t-distributions, in place of normal approximations, is not as crucial as you might be led to believe from a reading of some text books. The big differences between the exact t-distribution and the normal approximation arise only

for quite small sample sizes. Moreover the derivation of the appealing exactness of the t-distribution holds only in the idealized setting of perfect normality.

In situations where normality of the data is just an approximation, and where sample sizes are not extremely small, I feel it is unwise to fuss about the supposed precision of t-distributions.

3. Matched pairs

In his famous paper, Gosset illustrated the use of his discovery by reference to a set of data on the number of hours of additional sleep gained by patients when treated with the dextro- and laevo- forms (optical isomers) of the drug hyoscyamine hydrobromide.

patient	dextro	laevo	difference
1	0.7	1.9	1.2
2	-1.6	0.8	2.4
3	-0.2	1.1	1.3
4	-1.2	0.1	1.3
5	-0.1	-0.1	0
6	3.4	4.4	1
7	3.7	5.5	1.8
8	0.8	1.6	0.8
9	0	4.6	4.6
10	2	3.4	1.4

For each patient, one would expect some dependence between the reponses to the two isomers. It would be unwise to treat the dextro- and laevo- values for a single patient as independent. However, randomization lets us act as if the differences in the effects correspond to 10 independent observations, $X_1, \ldots, X_{10}$, on some $N(\mu, \sigma)$ distribution. (I have not checked the original source of the data, so I do not know whether the order in which the treatments was applied was randomized, as it should have been.) If there were no real difference between the effects of the two isomers we would have $\mu = 0$. In that case, $\overline{X}$ would have a $N(0, \sigma^2/10)$ distribution, and $T = \overline{X}/\sqrt{s^2/10}$ would have a t-distribution on 9 degrees of freedom.

The observed differences have sample mean equal to 1.58 and sample standard deviation equal to 1.23. The observed value of T equals

$$1.58/(1.23/\sqrt{10}) \approx 4.06$$

Under the hypothesis that $\mu = 0$, there is a probability of about 0.0028 of obtaining a value for $|T|$ that is equal to 4.06 or larger. It is implausible that the differences in the effect are just random fluctuations. The laevo- isomer does seem to produce more sleep.

REMARK.   One of the observed differences, 4.6, is much larger than the others. The value of 0 for the dextro- isomer with patient 9 also seems a little suspicious. Without knowing more about the data set, I would be wary about discarding the values for patient 9. I would point out, however, that without patient 9, the sample mean drops to 1.24 and the sample standard error drops to 0.66. The single large value made the differences have a much larger spread. The value of T increases to 5.66, pushing the p-value (for a t-distribution on 8 degrees of freedom) down to 0.0005.

4. Comparison of two sample means

Suppose we have two independent sets of observations, $X_1, \ldots, X_m$ from a $N(\mu_X, \sigma_X)$, and $Y_1, \ldots, Y_n$ from a $N(\mu_Y, \sigma_Y)$. How should we test the hypothesis that $\mu_X = \mu_Y$?

The two sample means are independent, with $\overline{X}$ distributed $N(\mu_X, \sigma_X/\sqrt{m})$ and $\overline{Y}$ distributed $N(\mu_Y, \sigma_Y/\sqrt{n})$. The difference $\overline{X} - \overline{Y}$ is distributed $N(\mu, \tau)$, with

$$\mu = \mu_X - \mu_Y \qquad \text{and} \qquad \tau^2 = \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}$$

If σ_X and σ_y were known, we could compare $(\overline{X} - \overline{Y})/\tau$ with a standard normal, to test whether $\mu_X = \mu_Y$.

If the variances were unknown, we could estimate τ^2 by

$$\widehat{\tau}^2 = \frac{s_X^2}{m} + \frac{s_Y^2}{n}.$$

The statistic $T = (\overline{X} - \overline{Y})/\widehat{\tau}$ would have approximately a standard normal distribution if $\mu_X = \mu_Y$. In general, it does not have exactly a t-distribution, even under the exact normality assumption. M&M pages 540–550 describe various ways to refine the normal approximation—things like using a t-distribution with a conservatively low choice of degrees of freedom, or using a t-distribution with estimated degrees of freedom. For reasonable sample sizes I tend to ignore such refinements. For example, I would have gone with the normal approximation in Example 7.16; I find the discussion of t-distributions in that example rather irrelevant (not to mention the unfortunate typo in the last line).

Pooling

If we know, or are prepared to assume, that $\sigma_X = \sigma_Y = \sigma^2$ then it is possible to recover an exact t-distribution for T under the null hypothesis $\mu_X = \mu_Y$, if we use a *pooled estimator* for the variance σ^2,

$$\widehat{\sigma}_{\text{pooled}}^2 = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2},$$

and

$$\widehat{\tau}_{\text{pooled}}^2 = \widehat{\sigma}_{\text{pooled}}^2 \times \left(\frac{1}{m} + \frac{1}{n}\right)$$

instead of $\widehat{\tau}^2$. If $\mu_X = \mu_Y$, then

$$(\overline{X} - \overline{Y})/\widehat{\tau}_{\text{pooled}}$$

has a t-distribution on $m + n - 2$ degrees of freedom.

Read M&M 550–554 for more about the risks of pooling.

5.   Summary

For testing hypotheses about means of normal distributions (or for setting confidence intervals), we have often to estimate unknown variances. With precisely normal data, the effect of estimation can be precisely determined in some situations. The consequence for those situations is that we have to adjust various constants, replacing values from the standard normal distribution by values from a t-distribution, with an appropriate number of degrees of freedom. When sample sizes are not very small, the adjustments tend to be small.

You should use the procedures based on t-distributions, in programs like Minitab, when appropriate. But you should not get carried away with thought that you are carrying out "exact inferences"—assumptions of normality are usually just convenient approximations.